Algorithmic Gatekeepers: The Human Rights Impacts of LLM Content Moderation

III. Right to Privacy

European Center for
Not-for-Profit Law

Acknowledgements:

Author: Marlena Wisniak, ECNL.

We extend special thanks to Isabelle Anzabi, who contributed significantly during the early stages of the research process. We express gratitude to the ECNL team—Karolina Iwanska, Vanja Skoric, and Francesca Fanucci—for their thoughtful review and feedback on the report.

Valuable input and review were provided by Evani Radiya-Dixit from the American Civil Liberties Union (ACLU); Lindsey Andersen from Business for Social Responsibility (BSR); Mona Elswah and Aliya Bhatia from the Center for Democracy and Technology (CDT); independent researcher and policy expert Luca Belli; and Roya Pakzad from Taraaz.

Insightful contributions through interviews and consultations came from representatives of Meta's Human Rights Team, the Policy and Safety Machine Learning Teams at Discord, and the Research Team at Jigsaw.

We extend our sincere gratitude to everyone who generously contributed their invaluable time, insights, and expertise to the preparation of this report. Your thoughtfulness and creativity have greatly enriched the quality and depth of our findings. We thank Betsy Popken of the UC Berkeley Human Rights Center; Corynne McSherry from the Electronic Frontier Foundation (EFF); Daniel Leufer and Eliska Pirkova from Access Now; Dave Willner of Stanford University; Dunstan Alison Hope; Jonathan Stray from UC Berkeley; Justin Hendrix of Tech Policy Press; Mike Masnick of Techdirt; Paul Barrett from New York University; Sabina Nong of Stanford University; Tarunima Prabhakar from Tattle; and Vladimir Cortes.

Design for the publication was created by Sushruta Kokkula and Andrea Judit Tóth. The illustrations featured in the report are based on the work of Balázs Milánik, Rozalina Burkova (The Greats) and Daniela Yankova (The Greats).

We thank the Omidyar Network for their generous support.

April 2025

Table of contents

Applicability of international human rights law for AI governance

International human rights law, grounded in instruments like the Universal Declaration of Human Rights (UDHR), the International Covenant on Civil and Political Rights (ICCPR), and the International Covenant on Economic, Social and Cultural Rights (ICESCR), provides a tested and globally recognised framework for assessing the potential risks and benefits of AI systems and content moderation—and enables a right to remedy. Human rights principles recognise inalienable rights, such as privacy, non-discrimination, freedom of expression, and freedom of peaceful assembly and association, which must be protected from undue interference. While these protections were historically focused on government obligations, the UN Guiding Principles on Business and Human Rights (UNGPs) have established that businesses—including AI companies— also have a responsibility to respect and uphold human rights.

As AI governance frameworks proliferate, many companies rely on ethics-based or trust and safety-driven approaches to responsible AI. While these frameworks often emphasise fairness, accountability, and harm mitigation, they typically lack consistency, international legitimacy, and are voluntary. By contrast, a human rights-based approach, legally binding for States, offers a universal, internationally recognised, and adaptable framework that applies across jurisdictions and industries and provide a right to remedy.

Given that AI-driven content moderation impacts human rights, integrating these principles into AI development, use, and governance can help AI companies navigate trade-offs and mitigate harm. Ultimately, it will help them protect and promote human rights in their products, services, and activities. International human rights also serve as a common baseline that enables meaningful collaboration between AI developers, deployers, regulators, and civil society, making them an essential foundation for evaluating and addressing risks in generative AI and developing rights-respecting products.

This report aims to highlight the key human rights impacts of using LLMs for content moderation, with a focus on core civic freedoms. While it doesn’t follow the methodology of a human rights impact assessment (HRIAs) under the UNGPs or a fundamental rights impact assessment (FRIAs) under the DSA or EU AI Act, our goal is to surface potential positive and negative impacts on a sector-wide level, to guide future HRIAs and FRIAs carried out by AI developers and deployers.

Right to Privacy

Legal basis

The right to privacy is protected under Article 17 ICCPR, whereas “No one shall be subjected to arbitrary or unlawful interference with his privacy, family, home or correspondence” and “Everyone has the right to the protection of the law against such interference or attacks.”[81]

In the EU, the EU General Data Protection Regulation (GDPR)[1] grants users rights to maintain control over their data. Companies that process data must have a valid legal basis, minimise data collection, and ensure that the data is processed solely for a specified, limited purpose. Moreover, the right to be informed (Articles 13 and 14) requires entities to provide transparent details about data processing, including the use of AI or LLMs. The right of access (Article 15) allows users to view their personal data. In the EU, privacy and data protection are not absolute and can be restricted under specific conditions as outlined in the EU Charter of Fundamental Rights. These rights must often be balanced with other fundamental EU values, human rights, or societal interests, such as freedom of expression, freedom of the press, and access to information. This balance ensures that privacy and data protection are upheld while accommodating other essential rights and public or private interests.[2]

While non-binding, the U.S. NIST AI Risk Management Framework warns that “the use of personal data for [generative AI] training raises risks to widely accepted privacy principles, including to transparency, individual participation (including consent), and purpose specification.”[82]

Inadequate data protection and lack of consent

In a 2024 paper, scholars from Stanford Human-Centered Artificial Intelligence cautioned that “AI systems are so data-hungry and intransparent that we have even less control over what information about us is collected, what it is used for, and how we might correct or remove such personal information. Today, it is basically impossible for people using online products or services to escape systematic digital surveillance across most facets of life—and AI may make matters even worse.[83]

A key issue is LLMs’ processing of data scraped from online sources without individuals’ knowledge and implication for cross-border data protection.[3] Indeed, the European Data Protection Supervisor warned that most of the data used to train advanced LLMs comes from publicly available internet sources, such as the Common Crawl dataset, which includes data from billions of web pages.[4] These datasets may

[1] General Data Protection Regulation (GDPR). (n.d.). *GDPR-Info.eu*. Retrieved February 27, 2025, from https://gdpr-info.eu/

[2] European Data Protection Supervisor. (n.d.). *Data protection*. European Data Protection Supervisor. Retrieved from https://www.edps.europa.eu/data-protection/data-protection_en

[3] Ibid.

[4] Lareo, X. (n.d.). *Large language models (LLM)*. European Data Protection Supervisor (EDPS). Retrieved from https://www.edps.europa.eu/data-protection/technology-monitoring/techsonar/large-language-models-llm_en

contain personal information about public figures and private individuals, which could be accurate, inaccurate, or outright misinformation. Managing the data protection risks associated with such datasets is highly complex.[5] Additionally, without proper safeguards, LLM outputs could inadvertently expose sensitive or private information from the training data, leading to potential data breaches.[6] Monitoring communication on nonencrypted platforms also undermines the expectation of privacy .[7]

When it comes to content moderation, consent and purpose limitation remain a critical yet often overlooked aspect of data collection, especially for training generative AI models. Many users are unaware that their data made available for other purposes—often gathered through web scraping, social media, or even public interest applications like language preservation efforts[8]—is being used for AI model training. Without clear mechanisms to ensure informed consent, specifically for training models, individuals may unknowingly give up their voice, language, or cultural expressions for commercial purposes, compromising their dignity and cultural rights. This lack of transparency undermines trust and raises ethical and human rights concerns about the commercialisation of personal and community data.

Indeed, LLM moderation relies on large-scale data, which may include personal information, such as names, addresses, or contact details, sometimes unintentionally scraped from platforms like social media.[9] Such data not only risks exposing platforms to breaches and hacks but can also be exploited through "model inversion," where models are hacked to extract copies of the data on which they were trained.[10] Moreover, users themselves might input sensitive details into AI tools without understanding that this data can be retained, aggregated, or even sold without their explicit consent. For instance, Instagram's changes to its Terms of Service, allowing data to be used for AI training unless users opt out, was successfully challenged in the EU.[11]

The large-scale collection and use of data for training LLMs, including for content moderation, creates unique privacy risks. Training datasets often include exchanges from public platforms, inadvertently capturing private and sensitive information. This raises concerns about the potential for AI outputs to leak such details to other users.[12] Additionally, the practice of using data for reinforcement learning and model improvement for content moderation blurs the lines of consent, turning everyday interactions into business opportunities without users' explicit approval.[13] These practices amplify vulnerabilities to privacy violations and underscore the need for stronger safeguards, including enabling privacy by design and as a default, as consistent with Article 25 GDPR.

It's finally important to note that there is an inherent tension between the right to privacy and the right to safety, as seen with traditional machine learning. To develop sufficiently effective automated moderation systems, models need to be trained or fine-tuned on platform-specific examples. However, some platforms have privacy

[5] Ibid.,

[6] European Data Protection Supervisor. (2024, January 17). *Large language models (LLMs)*. European Data Protection Supervisor. Retrieved from https://www.edps.europa.eu/data-protection/technology-monitoring/techsonar/large-language-models-llm_en

[7] Aliya Bhatia, personal communication, August 1, 2024.

[8] Ibid.,

[9] Office of the United Nations High Commissioner for Human Rights (OHCHR). (n.d.). *B-Tech right to privacy*. Retrieved from https://www.ohchr.org/en/b-tech-right-privacy

[10] Ibid.,

[11] Data Protection Commission (DPC). (2024, June 14). *The DPC's engagement with Meta on AI*. Retrieved from https://www.dataprotection.ie/en/news-media/latest-news/dpcs-engagement-meta-ai.

[12] Nicholas, G., & Bhatia, A. (2023, May). *Lost in Translation: Large Language Models in Non-English Content Analysis* (p. 30). Center for Democracy & Technology. https://cdt.org/wp-content/uploads/2023/05/non-en-content-analysis-primer-051223-1203.pdf.

[13] Aliya Bhatia, personal communication, August 1, 2024.

policies that prohibit collecting certain types of user data necessary to build functional models, making it impossible to implement automation in those contexts.

Inferring sensitive information

Recent advancements in LLMs have raised new privacy concerns, particularly around their ability to infer sensitive personal information from text. A 2024 study found that LLMs can accurately deduce attributes like location, income, and gender from real-world data, achieving high accuracy at a fraction of the cost and time required by humans (up to 85% top-1 and 95% top-3 accuracy at a lower cost of 100x and shorter time of 240x comparted to what is required by humans).[14] As more people interact with AI-powered chatbots, there's an increasing risk that these models could extract private data through casual questions. Moreover, researchers concluded that common privacy protections like text anonymisation and model alignment have proven to be ineffective, highlighting the need for stronger safeguards.

Google researchers themselves cautioned that for generative AI, "in addition to revealing sensitive information in training data, models may be able to correctly infer PII or sensitive data that was not in their training data nor disclosed by the user by stitching together information from disparate sources. These inferences can have a negative impact on an individual even if the inferences are not accurate (e.g., confabulations), and especially if they reveal information that the individual considers sensitive or that is used to disadvantage or harm them."[15] Under the GDPR, inferences—including false ones—are considered personal data. This means that all relevant obligations and rights apply, including individuals' right to access this data.

Personalised content moderation

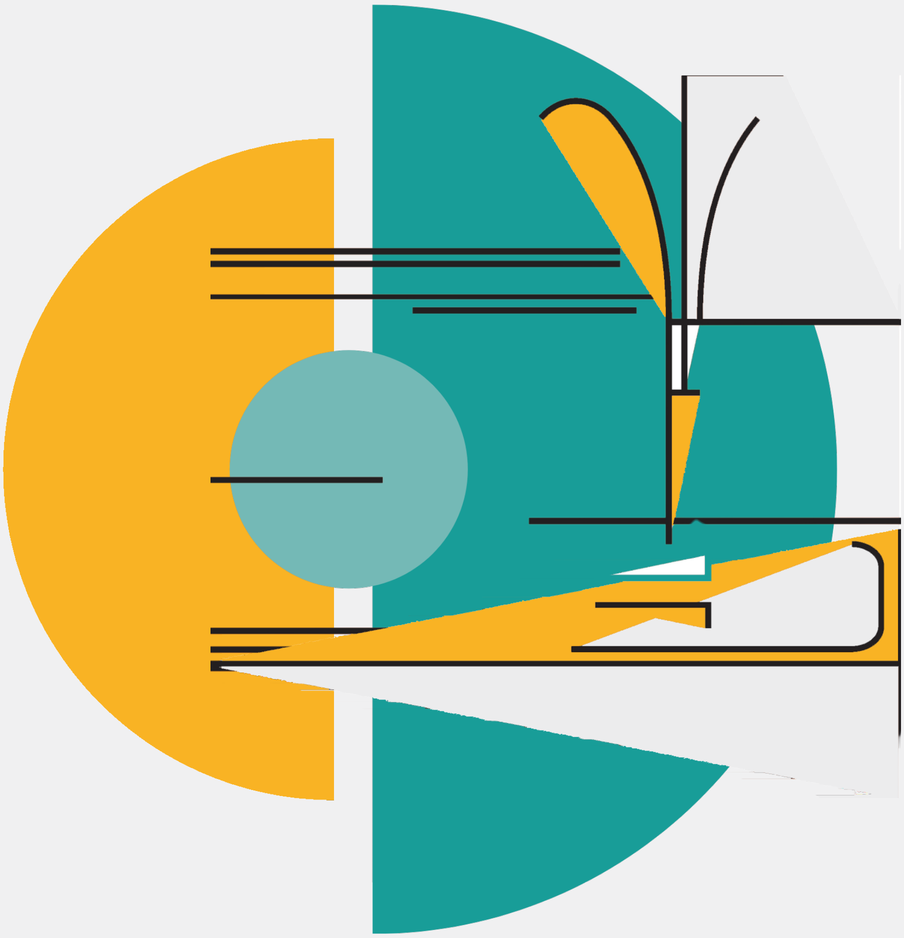
Generative AI and personalised content moderation offer the potential to enhance user experience, but they also come with considerable risks to privacy. To deliver highly personalised content or targeted advertisements, these systems rely on vast amounts of user data, incentivising companies to gather ever more personal information.[28] For instance, social media platforms and other companies implementing personalised content moderation may require users to share extensive details, such as interaction patterns, to improve personalisation. However, users are left in the dark about how much data is collected, how it is stored, and whether it is shared with third parties, including governments. This relentless data collection, often opaque and without meaningful user consent, directly threatens individuals' right to privacy.

As mentioned above, generative AI systems, particularly those embedded in recommender and content moderation models, can infer sensitive demographic information—such as age, gender, and race—even when it is not explicitly provided. This capability emerges as algorithms detect patterns in user behavior and reconstruct sensitive attributes. While platforms may not intentionally design these systems for demographic targeting, the emergent effects of personalised recommendations can lead to such outcomes. For example, viewing history and interaction data can inadvertently reveal details like race or socioeconomic status.[29] This information can then be used to create targeted advertisements or content, often without users' knowledge or consent.

[14] Staab, R., Vero, M., Balunović, M., & Vechev, M. (2024, May 6). Beyond memorization: Violating privacy via inference with large language models (Version 2).arXiv. https://arxiv.org/abs/2310.07298v2

[15] Pearce, A. Jiang, E. (2020, December). Data leak: Exploring the impact of data privacy breaches. PAIR Google. Retrieved from https://pair.withgoogle.com/explorables/data-leak/

Beyond personalisation, the data collected and inferred through LLM-powered systems also increases the risk of profiling. By combining explicit and inferred data, platforms can create highly detailed user profiles that can be used for commercial purposes to influence users' behaviours and which content they engage with. They might also be shared with governments or other third parties, including advertisers. Such profiling has significant implications when governments, particularly those with authoritarian practices, leverage this data for surveillance or repression, or when political parties use this data to micro target their campaigns.[30] What's more, when companies are not transparent about how long data is stored or who has access to it, the risk of misuse increases, especially for law enforcement. This is particularly concerning in cases where platforms rely on LLMs to improve interaction-based content moderation, as the data required for such systems inherently grows with each user interaction.



Government surveillance

Surveillance of online content poses a serious threat to the right to privacy.[16] Historically, government and law enforcement agencies have monitored social media platforms to track individuals or groups of interest, often targeting marginalised communities based on factors like race, religion, or political beliefs.[17] The specific technologies and methods employed in such surveillance are likely to remain obscure, justified under the guise of 'security' concerns. Yet LLMs' text classification capabilities could provide new opportunities for mass surveillance, raising concerns about their potential misuse in monitoring and analyzing online content.[18] Little is known about governments' investment around the world in LLM tools to localise these technologies for domestic uses, which raises concerns about government interference and control.

Furthermore, generative AI models trained on data scraped from the internet might inadvertently retain personal information about individuals, including details about their relationships with family and friends.[19] Governments, especially those with authoritarian practices, could thus misuse user data and LLM tools for harmful purposes. Indeed, the Digital Trust & Safety Partnership (DTSP), an industry group formed by leading digital platforms, acknowledged that "GenAI could enable new levels of government-compelled surveillance. Governments could theoretically compel companies to use genAI to enforce local content restrictions by law, which would be especially problematic for users in authoritarian contexts."[20]

In the hands of governments, these tools can become mechanisms of repression and control. Governments with authoritarian practices have already exploited content moderation systems to monitor and silence dissent. LLMs exacerbate these risks due to their advanced capacity to infer personal characteristics, such as political beliefs or affiliations, from user-generated content. This technology can be weaponised to track journalists, human rights defenders, and marginalised groups, as seen in Iran, where inferred profiling has been used to track and criminalise people in the past, especially women[21] and civic space actors.[22]

The danger lies not only in monitoring existing content but also in predicting and controlling future behaviours. For instance, LLMs can flag content likely to gain traction and suppress its reach before it garners attention. In authoritarian states like China, governments already use machine learning to monitor and censor content based on predefined keywords, but LLMs introduce a new level of precision and scalability.

[16] Jeroudi, L. (2023, June). *Surveillance and human rights: Background paper.* Global Coalition on Human Rights and Digital Surveillance (GCHRAGD). Retrieved from https://gchragd.org/wp-content/uploads/2023/06/GCHRAGD-SURVEILLANCE-AND-HUMAN-RIGHTS-background-paper.pdf

[17] Levinson-Waldman, R., Panduranga, H., & Patel, F. (2022, January 7). *Social media surveillance by the U.S. government*. Brennan Center for Justice. Retrieved from https://www.brennancenter.org/our-work/research-reports/social-media-surveillance-us-government

[18] Ibid.,

[19] Miller, K. (2024, March 18). *Privacy in an AI era: How do we protect our personal information?* Stanford HAI. Retrieved from https://hai.stanford.edu/news/privacy-ai-era-how-do-we-protect-our-personal-information

[20] Digital Trust and Safety Partnership. (2024, September). *Best practices for AI and automation in trust and safety* (p. 41). Retrieved from https://dtspartnership.org/best-practices-for-ai-and-automation-in-trust-and-safety/

[21] Parent, D. (2025, March 24). *Drones, informers and apps: Iran intensifies surveillance on women to enforce hijab law. The Guardian.* Retrieved from https://www.theguardian.com/global-development/2025/mar/24/iran-police-women-surveillance-hijab-drones-dress-code-law.

[22] Miaan Group. (2024, June). *The internet in the Women, Life, Freedom era: Iran's progress in censorship and surveillance – and options for European policymakers.* Friedrich-Ebert-Stiftung. Retrieved from https://library.fes.de/pdf-files/international/21296.pdf.

Individuals associated with flagged content face severe offline consequences, including violations of their right to liberty and security through detention, torture, or killings. This confluence of predictive power and real-world sanctions creates a chilling effect, silencing dissent and fostering self-censorship on a massive scale, which is explored in the section on freedom of expression.

While open-source LLMs promise transparency and community-driven innovation, they also lower the barriers for misuse, especially by malicious actors. Governments with authoritarian practices, often equipped with significant infrastructure and state-controlled data, can adapt open-source content moderation models to target specific populations or behaviours.[23] This raises critical questions about who is using these models, how they are applying them, and for what purposes. For example, open-source content moderation models can easily be repurposed for surveillance, enabling governments to monitor LGBTQIA+ communities, political dissidents, or those seeking information on banned topics, such as abortion rights. The use of open-source models for content moderation also poses unique challenges in terms of accountability. Once a model is publicly released, its use cannot be easily regulated or monitored. Governments and other malicious actors can exploit these tools to conduct mass surveillance, mine social media for sensitive data, or even generate targeted disinformation campaigns to manipulate public opinion. With LLMs capable of analysing and summarising large datasets rapidly, the potential for widespread privacy violations grows exponentially.

[23] Sabina Nong, personal communication, September 13, 2024

Algorithmic Gatekeepers III. Right to Privacy

European Center for
Not-for-Profit Law